A Cloud-based System for Scraping Data From Amazon Product Reviews at Scale

Ryan Woodall
Jrw5074@uncw.edu

Douglas Kline
klined@uncw.edu
Information Systems

Ron Vetter
vetterr@uncw.edu
Computer Science

Minoo Modaresnezhad
modarsm@uncw.edu
Information Systems

University of North Carolina Wilmington
Wilmington, NC 28403

Abstract

Amazon product reviews can provide a rich source of data for natural language processing research. To support a related research project, we built a custom cloud-based system for obtaining Amazon product reviews. A third party cloud-based scraping service automatically retrieved scraping jobs, then notified Azure Data Factory through an Azure Function. Raw scraping data was then transferred in batches to Azure Data Lake Storage, then custom SQL transformed the data for convenient query from an Azure SQL database. The system was used to obtain 17,962 product reviews and produce data sets in several formats. This paper fully describes the system, and offers lessons learned from the experience.

Keywords: Data Pipeline, Cloud, Amazon Reviews, Big Data, Azure.

An updated version of this manuscript may be found at https://jisar.org